PECAN: a novel 16S rRNA gene sequence non-clustering based taxonomic assignment tool

Johanna B. Holm[1], Pawel Gajer[1], Jacques Ravel[1]

[1]Institute for Genome Science, University of Maryland Baltimore School of Medicine

PECAN
Per sEquenCe tAxonomic assigNer

Background

Clustering of sequences into Operational Taxonomic Units (OTUs) has become a mainstream approach to facilitate taxonomic classification of large numbers of 16S rRNA gene sequences. This is partly due to the high computational requirements for processing each sequence in increasingly large datasets. A primary focus of the field has been development and improvement of OTU-based sequence clustering methods that rely on distances between each pair of sequences in a dataset. Following OTU-based clustering, representative sequences are commonly classified using tools such as the RDP Naïve Bayesian Classifier, and the resulting classification transitively assigned to all sequences comprising that OTU. However, problems with this strategy exist[1]. Here, we present PECAN, a novel per sequence taxonomic assigner which quickly and accurately classifies millions of 16S rRNA gene sequences. PECAN relies on higher order Markov Chain models built from a user-specified set of reference sequences. These models are used to estimate the probability that a query sequence belongs to a particular taxonomic rank.

Methods & Usage

1. Align sequences - Mafft v 7.222
2. Produce phylogenetic tree - FastTree JC +CAT
3. Automated taxonomic annotation curation (Figure 1)

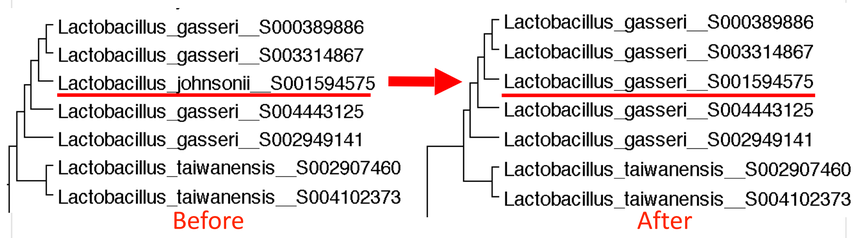

Figure 1. Example of taxonomic annotation curation

Automated reference sequence curation & automation is an ongoing project

4. Curation produces a taxonomic tree containing all sequences (Figure 2). For each species, genus, etc., build 8th order Markov Chain model (Table 1).

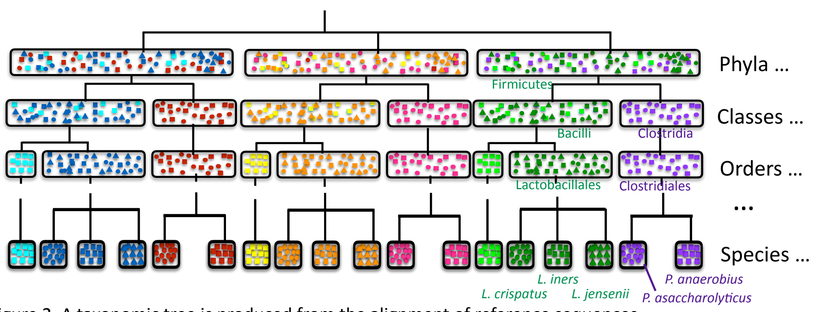

Figure 2. A taxonomic tree is produced from the alignment of reference sequences

Table 1. Markov Chain models of the probability of observing a particular base following a particular 7mer using *Lactobacillus* spp. as an example.

	A | 7mer$_a$	**C | 7mer$_a$**	**G | 7mer$_a$**	**T | 7mer$_a$**	**A | 7mer$_b$**	**C | 7mer$_b$**
L. iners	P (A | 7mer$_a$)	P (C | 7mer$_a$)	P (G | 7mer$_a$)	P (T | 7mer$_a$)	P (A | 7mer$_b$)	P (C | 7mer$_b$)
L. jensenii	P (A | 7mer$_a$)	P (C | 7mer$_a$)	P (G | 7mer$_a$)	P (T | 7mer$_a$)	P (A | 7mer$_b$)	P (C | 7mer$_b$)
L. crispatus	P (A | 7mer$_a$)	P (C | 7mer$_a$)	P (G | 7mer$_a$)	P (T | 7mer$_a$)	P (A | 7mer$_b$)	P (C | 7mer$_b$)
...	P (A | 7mer$_a$)	P (C | 7mer$_a$)	P (G | 7mer$_a$)	P (T | 7mer$_a$)	P (A | 7mer$_b$)	P (C | 7mer$_b$)
g_*Lactobacillus*	P (A | 7mer$_a$)	P (C | 7mer$_a$)	P (G | 7mer$_a$)	P (T | 7mer$_a$)	P (A | 7mer$_b$)	P (C | 7mer$_b$)
...	…	…	…	…	…	…
f_Lactobacillaceae	P (A | 7mer$_a$)	P (C | 7mer$_a$)	P (G | 7mer$_a$)	P (T | 7mer$_a$)	P (A | 7mer$_b$)	P (C | 7mer$_b$)
...	…	…	…	…	…	…
o_Lactobacillales	P (A | 7mer$_a$)	P (C | 7mer$_a$)	P (G | 7mer$_a$)	P (T | 7mer$_a$)	P (A | 7mer$_b$)	P (C | 7mer$_b$)
...	…	…	…	…	…	…
c_Bacilli	P (A | 7mer$_a$)	P (C | 7mer$_a$)	P (G | 7mer$_a$)	P (T | 7mer$_a$)	P (A | 7mer$_b$)	P (C | 7mer$_b$)
...	…	…	…	…	…	…
p_Firmicutes	P (A | 7mer$_a$)	P (C | 7mer$_a$)	P (G | 7mer$_a$)	P (T | 7mer$_a$)	P (A | 7mer$_b$)	P (C | 7mer$_b$)

"For *L. iners*, this is the probability of observing a C given the previous 7 nucleotides" (here, labeled as "7mer$_b$")

Methods & Usage (cont.)

5. The reference taxonomic tree is now represented by Markov Chain models (Figure 3).

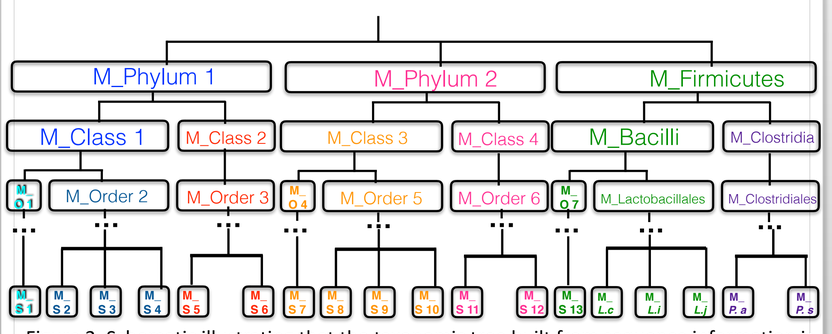

Figure 3. Schematic illustrating that the taxonomic tree built from sequence information is now represented by Markov Chain models annotated as "M_".

6. To classify a query sequence calculate its 8th order Markov chain model, and match model to those of reference tree. PECAN first determines the phylum, and then consider only those nodes within that phylum, speeding up classification (Figure 4).

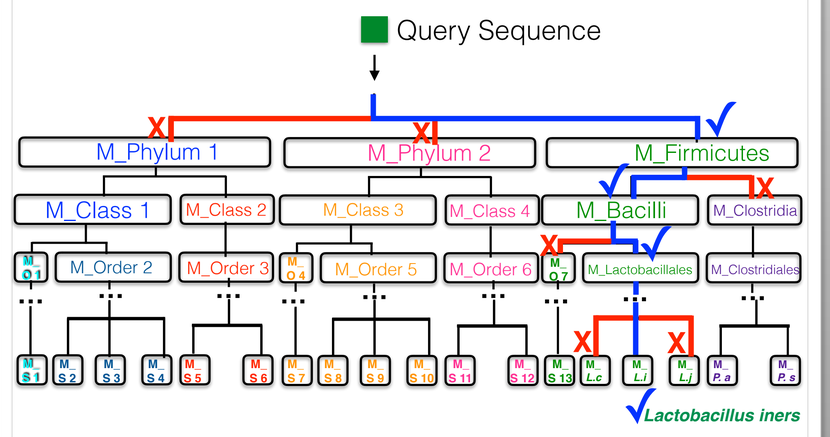

Figure 3. A query sequence is taxonomically assigned with PECAN. The 8th order Markov chain model of that sequence is matched with those of the reference tree starting at the phylum level. Groups considered are indicated by a red (mis-match) or blue (match) line. Here, the query sequence is classified as *Lactobacillus iners*.

7. Figure for distribution showing scoring threshold

1. For all reference sequences in group, list all 8-mers
2. Using all reference sequences in a group, calculate conditional probability of observing a base given 7mer preceding the base:

Results

Summary

PECAN:

- Requires less computational resources than other algorithms
- Is fast! Individually classifying over 1 million sequences per minute
- Removes the necessity and issues associated with OTU-based sequence clustering
- Is applicable to any microbiome dataset by customization of the reference sequences including

References

PECAN: https://github.com/pgajer/MCclassifier